浅谈全媒体内容库的内容挖掘与可视化

摘　要：随着全媒体内容库的建设，内容库中的资源数量呈指数增长，海量的信息出现使得用户感觉无所适从，很难从中寻找到真正需要的内容资源，出现了互联网行业所谓的“信息过载”现象。为了能够使用户更容易地利用和理解内容库中的多种、大量的媒体资源，更好地推进全媒体内容服务的特色化和个性化服务，本文分析了现阶段全媒体内容库在内容管理、内容服务，挖掘内容价值上出现的新需求，描述了各种媒体可用的分析挖掘技术，最终通过信息可视化技术将全媒体内容库中的内容直观地展示在用户面前，使用户更加容易理解和利用内容库的全媒体内容。

关键词：融合媒体；全媒体内容；智能化；内容挖掘；内容可视化
中图分类号：G206　　文献标识码：A
文章编号：1671-0134（2018）07-079-03　　DOI：10.19483/j.cnki.11-4653/n.2018.07.024

文 / 郭海 [1]　程大川 [2]

1. 全媒体内容服务的新需求

全媒体内容库是融合媒体平台的重要组成部分，通过全媒体内容库可实现跨媒体的内容资源管理，整合全台在线全媒体内容资源。通过构建全媒体内容库，还可以实现全台内容的统一检索、统一共享和快速调用，真正激活台内现有的媒资及各种业务系统内的媒体内容资源，为全媒体融合生产、全媒体指挥策划和内容运营提供内容支撑。当前，广电融合媒体平台建设如火如荼，在新的技术平台和业务架构下，媒体内容管理的对象、流程，以及提供内容服务的方式也随之变化。

用户需要提供更加丰富的内容发现手段，不仅仅是分类查找、全文搜索。在用户有明确目的查找内容的时候，要能使用户随时随地通过各种搜索手段获得准确的内容。在用户没有明确目标的时候，达到“想你所想”的内容响应，在服务形式上化被动为主动，将内容与用户需求相结合，为用户提供精准、贴合的内容分析服务。

原有内容再造，是传统媒体在内容上的核心竞争力，[1] 全媒体内容库需要为用户提供更多的内容可视化分析工具，让用户通过内容可视化工具，更多维度的去理解内容库中的内容，让内容的分析、内容挖掘更加简单、易用、直观，让用户进行交互式、可视化的内容探索。

2. 内容可视化技术研究与应用

全媒体内容库以大数据、人工智能等先进技术为支撑，进行媒体内容服务的创新，通过对全媒体内容进行标签提取，基于内容标签进行内容的深度分析、计算，实现对内容库媒体内容的挖掘，包括索引、关键词提取、自动摘要、分类、聚类、情感分析、关联计算等，对得到的分析计算结果进行解释和表示。最终，通过内容可视化的方式为用户提供更友好、更准确的内容服务。



2.1 全媒体内容的特征提取及计算

全媒体内容库中的内容种类包括文本、图片、音频、视音频等，全媒体内容包含多种维度的描述内容的特征，对于这些特征的提取，是分析、挖掘全媒体内容的基础。

2.1.1 文本特征提取

全媒体内容库通过自然语言处理技术，实现对文本内容的分析，包括词性分析、关键词提取、自动摘要、情感分析等，从而提取文本的相关特征。

2.1.2 图像特征提取

通过人工智能图像技术，将内容库中的图像内容自动生成相应的文字描述，描述的特征范围涵盖场景描述、物体分类、人物、地标、热词等，可以提取对不同维度、不同层次的图片内的语义特征信息。[2] 抽取相关的特征向量后，形成代表该图像的多维特征向量，通过向量计算，在特征向量空间中比较、分析各图像特征向量之间的距离或相似关系，完成对图像内容相关分析计算，从而让系统获得高层次的对图像的理解。

2.1.3 视频特征提取

通过人工智能的视频处理技术，对内容库中的视频进行智能分析，分析的视频内容包括视频内的语音、文字、人脸、物体、场景等多种维度内容，提取描述视频内容的特征信息。特征内容包括视频的分类、人物、语音识别文字、物体、字幕、标题、弹幕文字内容等。

2.1.4 音频特征提取

通过人工智能技术，对内容库中的音频进行智能分析。音频挖掘通常有两种方式：将音频中的语音识别成文字，再对文字信息进行特征提取；从音频中提取音乐

特征，例如音调、旋律，以及说话人的声纹特征，提取人物等。通过分析获得的音频特征，进行音频内容的相关内容分析计算。

2.2 基于标签的内容分析

所谓标签，既上述媒体内容的特征，通过一些标签描述、代表某一内容，这个内容可以是全媒体内容库中的任何内容，包括互联网汇聚、通联稿件、报题、选题、报道等。对比传统“分类”的概念，一个内容往往只属于几个固定的分类，但可以拥有许多个不同维度的内容描述标签。同时，“分类”的定义相对固定，数量有限，而标签则是数量无上限扩展，自由添加，并且标签具有时效性。全媒体内容库中的内容大爆炸给标签带来了用武之地，通过内容标签可以对全媒体内容库中的海量内容进行多维度的管理。

通过标签进行内容分析，让系统、用户可以更多维度、更深地理解内容，抽取全媒体内容的具有表意性、显著性的特征、特点，形成描述该全媒体内容的标签。

可以解决使用者对海量全媒体内容的“发现”的需求。内容库中，海量的全媒体内容，使用者要想快速、方便的找到自己需要的内容，通过内容标签是最方便、可靠地方式。对于全媒体内容库的使用者，标签能够有效地进行内容整理、组织和管理。

标签有助于用户挖掘全媒体内容库中的内容价值，通过人工智能技术挖掘得到不同维度的内容标签，使用户进行探索式内容挖掘成为可能，用户可以从不同角度通过可视化的方式，观察内容库中的内容，发掘更深度的内容价值。

2.3 多种媒体内容的可视化应用

对于全媒体内容库中的内容可视化，处理的数据类型涵盖文本、多维、视音频、时空数据等，采用的显示方法包括标准的2D/3D图表显示、图像化显示、列表显示、地图显示等。

多维数据可视化分析的目标是发现多维数据分布规律，寻求不同维度之间的逻辑关系，通常采用的显示方法包括曲线图、折线图、散点图、饼图、柱状图、雷达图、热力图等图表显示方法。

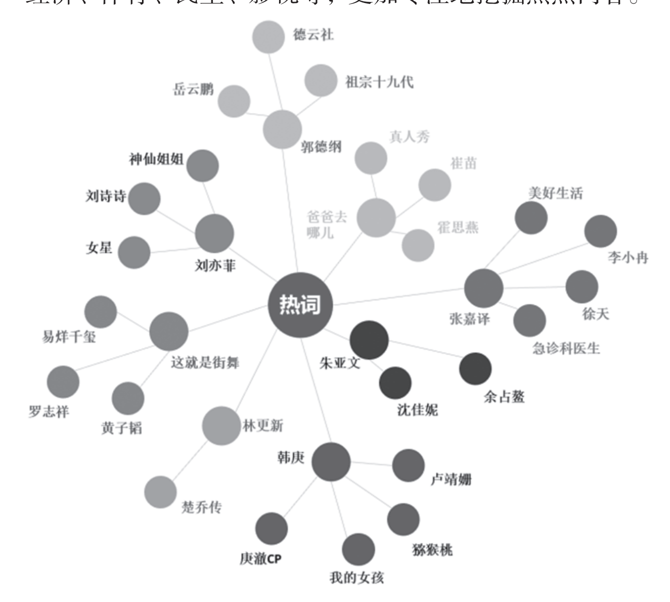
文本内容的可视化，常用的可视化技术是标签云，[3] 它直接抽取文本中的关键词并将其按照一定的顺序和规律整齐美观地呈现在屏幕上。关键词在文本中有分布的差异，有的出现的频率高，有的出现的频率低，有的重要性高，有的重要性低，可以利用字体的大小和颜色的醒目度反映文本中各个关键字的差异，越是重要、出现频率高的关键词可以采用较大、颜色较醒目的字体。

2.4 探索式可视化的内容分析

可视化的自助式内容探索工具，辅助用户通过可视化的方式分析、挖掘内容，产出对内容生产有价值的洞察。整个内容挖掘“可视化”的过程，用户根据需求简单进行拖拽式、交互式操作即可完成，多种展示形式，秒级响应。让用户能够以最直观的方式发现一些内容背后潜在的相关性。

（1）内容分布分析工具，通过内容的属性、标签等，用户可以通过工具自助地进行内容分析、统计，提供多种可视化组件，使用户以最直观的方式了解内容库中的内容。

（2）热点内容挖掘工具，通过可视化工具可以直观地发现当前热点，再通过热点进行下钻，分析、挖掘与热点相关的内容。可以通过选择不同领域，包括时政、经济、体育、民生、影视等，更加专注地挖掘热点内容。



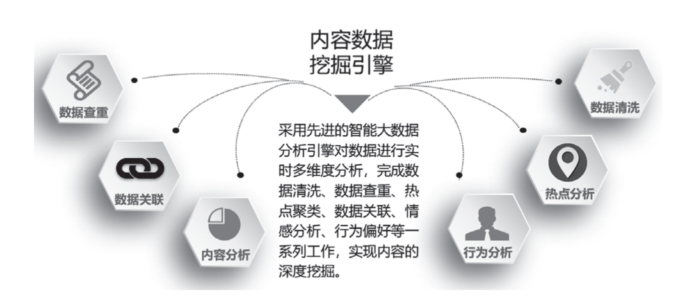
（3）主通过对海量内容的智能挖掘关联，自动聚合生成事件专题、人物专题、自定义专题等。通过可视化的对内容聚类分析结果的呈现，帮助用户挖掘潜在的内容，为选题决策、内容生产提供智能支持。

（4）词云分析工具，以所选范围内的内容相关的标签数据为分析基础，以词云的方式对标签数据进行展示，用最直观的方式为业务人员展示舆情关键词、新闻关键词、评论关键词等。

（5）情感分析工具，通过对评论、舆情内容的智能分析，利用先进的深度学习技术，实现内容的精准挖掘，针对媒体领域特点进行情感正负面判断，提取与用户相关的高价值信息。[4]

3. 全媒体内容库构建的核心模块

全媒体内容库致力于打造出大容量、多种类、可学习、可交互的智能化内容管理引擎，通过全媒体内容服务平台，可以充分吸收多种渠道海量汇聚的内容，经过内容整理后，实现内容的精准查询、关联检索、可视化分析等。



同时，还可以根据对自有内容的数据挖掘，通过建模，自动形成主题事件库、知识库等面向业务的辅助决策、辅助生产、辅助发布的内容池。核心模块包括以下几方面。

3.1 内容处理引擎

针对不同类型、不同来源的内容，处理引擎对内容进行自动化处理，进行结构化，内容处理引擎能力包括：

内容筛选：通过分析内容元数据、文本信息，以及系统配置信息，设置内容的重要级别、保密级别。

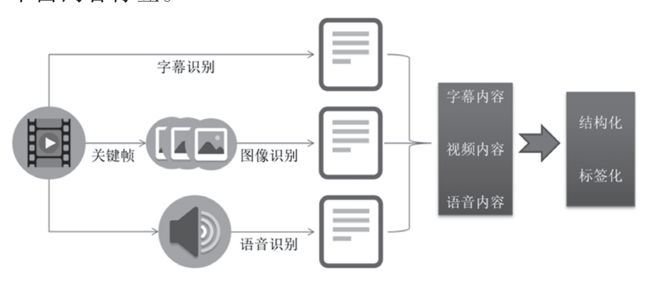
内容过滤：内容的重复过滤、垃圾信息过滤、广告过滤，同时对视音频素材也需要具有过滤功能，对重复上传的视音频避免重复入库。

内容审核：基于敏感词及特征库，过滤检测内容的文本、图片及视频。自动过滤汇聚素材中的敏感内容，并将包含敏感词的素材放入待发布区，由人工二次处理。

字幕检测与识别：从视频中检测到是否有字幕。字幕识别，对有字幕的视频把字幕转换成文字。

人脸检测：检测是否是某个特定人的脸，检测是否含有人脸。

视频标签提取：识别视频中的场景、人物、风景、建筑、生活物品等，支持不同维度层次的图像语义信息提取，丰富内容标签。



多格式转换：可将多种音视图文素材转换为标准的制作格式、码率，供电视新闻生产网或新媒体调用。

3.2 内容挖掘引擎

通过人工智能技术对入库的海量内容的属性、文本、标签等进行内容挖掘，包括分类、聚类、自动关联、实体名提取、情感分析、标签提取等。

分类：通过分析内容元数据、文本信息、内容标签，以及分类配置信息，自动将入库内容进行分类。

聚类：根据编目信息、自动提取的标签信息，支持文件属性自动辨别分类，元数据分析分类，元数据自动关联，话题内容相似性聚类，可通过自动聚类技术自动聚焦一段时间内网上热点信息，可自定义需进行自动聚类运算的素材来源和类别。

自动关联：新闻稿件相关内容自动关联，自动关联相关、相似的多媒体素材，形成新闻素材集合，便于编辑制作人员有针对性地挑选采用，可自定义需进行自动关联的来源，可灵活配置自动关联分析灵敏度。

实体名提取，对内容进行领域内的实体名提取，包括栏目名、节目名、主演、主持、导演等领域内实体名；

情感分析，针对舆情、评论，进行情感分析。

标签提取：通过分析内容的元数据及文本信息，自动提取内容的关键词，形成内容的标签。

3.3 可视化渲染引擎

通过灵活使用 HTML5 技术，适配不同的展现模式，同时运用 CSS3 的动画特性，结合媒体内容的自身特点，以更生动、更友好的形式，实时呈现隐藏在庞杂媒体内容背后的规律、联系。

总结

以全媒体内容为基础的媒体融合业务，对内容保存和使用需求不再只是以素材和节目为核心，也不再以人工编目和结构化的数据保存为主要手段，而是扩展到面向全媒体业务，涵盖素材、节目、电视稿件、两微内容、H5 页面等多种内容形态，同时，对内容的编目也以自动化的数据提取、智能编目、非结构化的原始数据保存为主要手段，重视对原始内容数据的持续挖掘。[5]



以大数据、人工智能技术为基础，重新梳理媒体内容服务的各个环节。基于标签的内容分析挖掘，充分发挥内容的最大价值，最终实现全台内容包括媒资、制作，以及电视媒体、广播媒体、新媒体等多种业务体系内容的统一检索和使用，实现为融合媒体各种业务的内容支撑。媒

参考文献

[1] 李正平，温序铭 . 全媒体时代的电视台内容服务——全媒体媒资建设 [J]. 现代电视技术，2012（02）：142-143.

[2] 陈东 . 数字图书馆中多媒体数据挖掘技术研究 [J]. 图书馆学研究，2004（05）.

[3] 唐家渝，刘知远，孙茂松 . 文本可视化研究综述 [J]. 计算机辅助设计与图形学学报，2013（03）：273-282.

[4] 张瑞 .SMS 网络舆情信息监控系统的设计与实现 [J]. 现代情报，2012（03）：68-71.

[5] 张楠 . 数字媒体视阈下的新闻制作与传播 [J]. 新闻研究导刊，2016（03）：80.

（作者单位：1. 山西广播电视台；2. 新奥特（北京）视频技术有限公司）